EPSRC
Engineering and Physical Sciences Research Council

InFoMM
Industrially Focused Mathematical Modelling

EPSRC Centre for Doctoral Training in Industrially Focused Mathematical Modelling

Spatio-temporal networks in supermarkets

Fabian Ying

UNIVERSITY OF OXFORD

TESCO

Contents

1. Introduction

Background

How do customers move and navigate within a supermarket? What is the best layout of a store to reduce congestion? Should shelves with promotional items be located in the middle or at the end of aisles? These and other questions are of considerable interest to retailers and marketing researchers. Until recently, however, only qualitative studies have been conducted and these studies are based predominantly on observations (by following customers around a store) and surveys.

Anonymized tracking of customer trolleys could help understand and reduce congestion within stores.

Thanks to recent technology, customer trolleys can now be tracked using radio frequency identification (RFID) chips, allowing large-scale, anonymized, and non-invasive tracking of customer journeys. This makes it possible to conduct more rigorous investigations of the key questions. As part of a trial for this trolley-tracking technology, Tesco has collected trolley location data in one of its stores and provided us with six days' worth of data.

Our approach in this project is to build a simple mathematical model for customer movements within a store. We rely on models for human mobility which have been a major research topic during the last decade thanks to the recent availability and abundance of human and animal mobility data. Models with simple assumptions about individual behaviour that can reproduce the overall observed behaviour are useful to explain the underlying dynamics, as well as to predict them.

2. Overview

Total dwell time:
Total amount of time that customers spends in a location/zone
Visits:
Number of times that customers entered a location/zone

A *network* is a set of objects (called *nodes*) that are connected by *edges*.

Our aim is to develop a model for customer journeys that can accurately predict quantities of interest such as total dwell time and number of visits of each region in the store (see box for definition). After representing a store as a network (see Figure 1), we apply our model to this network and we compare quantities of interest between model and data.

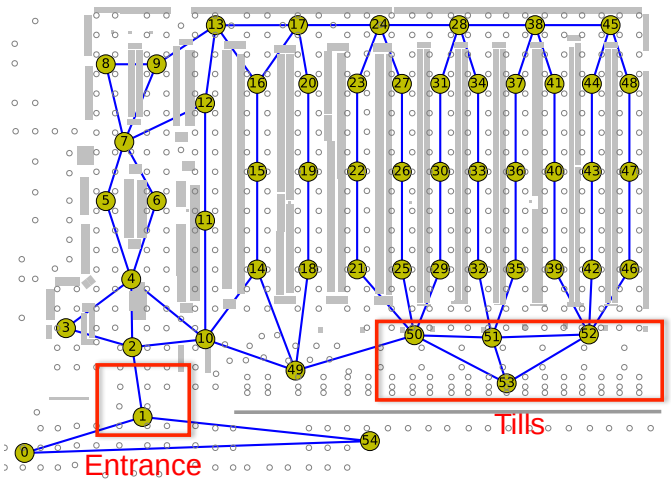

Figure 1: Layout of the store represented as a network. The nodes (circles) represent the zones within the store. Adjacent zones are connected by edges (blue lines). The entrance is in zone 1, and the tills are in zone 50 – 53.

Trolley locations are recorded every 2 seconds with a precision of about 4 square meters

3. Data

The trolleys are tracked using RFID chips, which record their approximate location every two seconds. There are 720 locations distributed around the store (see Figure 2) and the data records the coordinates of the locations, to which the trolley was closest at each timestamp. This gives a resolution of about 4 square metres.

In addition to the trolley location data, Tesco provided us with sales and product location data, which we use to estimate the number of items sold in each zone.

Identifying customer paths from raw data

The trolley location data do not distinguish when a trolley is used by a staff member or a customer, so we need to identify which paths are those of customers. In our preliminary method, we formulated the following conditions that customer paths must satisfy.

1. The path starts in the zones near the entrance area (zones 1, 2, 3);
2. The path ends at the area behind the till (zone 53);
3. The path does not enter the warehouse (zone 55) at any time;
4. Prior to the start of the path, the trolley must come from outside (zone 0 or 54);
5. Prior to entering zone 53, the trolley must pass through the queue area (zone 50, 51, 52 or 53);
6. Paths must have a journey time of at least 5 minutes.

Using these conditions, we identified 324 likely customer paths. Surprisingly, this makes up only 3.6% of the data (the majority of the rest is idle trolley data) and we discarded the rest. In particular, we discarded all paths that finished in less than 5 minutes (condition 6), which make up a significant proportion of the paths. We hope a store visit will help us give insight into distinguishing short staff path from short customer paths.

Total dwell time

In Figure 2, we show the total dwell time (that is, the time the customer lingers with their trolley) for each location. We observe that customers spend more time in the left part of the store than in the right part; this is not surprising, given that more popular items, such as produce, bread and bakery are on the left side, and less popular items are on the right side of the store.

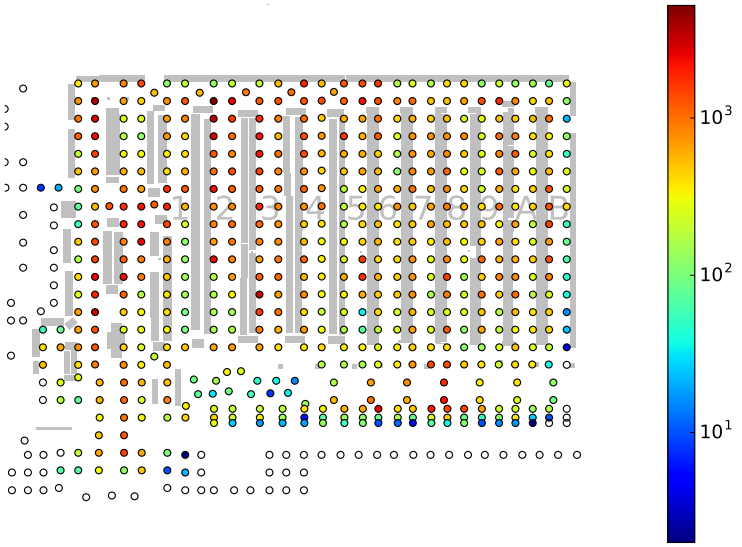

Figure 2: Visual map of total dwell time (in seconds) for each location inside a store. Red is higher, blue is lower. Customers spend more time in the left part of the store, which generally stocks food products, than in the right part, which stocks non-food merchandise.

In a *biased random walk* the customer makes random steps, but with tendency in a direction towards the zones that contain their desired items.

Model parameters

p = probability of taking a detour

L = shopping list

ω_i = attraction values of zone i

4. Mathematical model

We formulate a simple mathematical model to describe customer journeys which is based on biased random walks on a network. The store is represented as a network by dividing the store into *zones* (which are the nodes of the network) and connecting neighbouring zones by edges. We divide the store into 56 zones of similar sizes and create a network of the store (see Figure 1).

Our model is built around the key parameters p, L and ω_i. We assume that a customer starts in zone 1 (i.e., at the entrance) with a *shopping list* L, which we model as a list of zones that they intend to visit. The list L can be generated, for example, based on actual purchase histories or sales data. At each time, the customer either walks towards the closest zone that contains an item in their shopping list or takes a detour step to explore items in a neighbouring zone or their current zone.

The probability that the customer takes a detour step and goes at random to either a neighboring zone or stays at current zone is denoted by p. When the customer takes a detour step, the next zone is chosen with probability proportional to its attraction value ω. Zones with higher attraction values are thus visited more often during detour steps. Otherwise, if no detour is taken, the customer moves a step towards the nearest item (or zone) in L. That is, they choose the neighbour from which it requires the least number of step to reach a zone in L. Whenever a zone in L is visited, this zone is removed from L (because all desired items in that zone were picked up). When L is empty (as all items on the shopping list have been picked up), the customer heads towards the tills (zone 53 in our graph). However, on the way to the tills, the customer may still take detour steps.

Comments

- Our model assumes that customers only look forward one step at a time and take only local, one-step detours.
- One can interpret customers with low detour probability p as goal-oriented shoppers or shoppers familiar with a store. Customers with high detour probability p are customers with more exploratory behaviour.
- In our simulations, we generate the shopping list L randomly, with each zone having the same probability of being selected. For simplicity (and for now), we let each customer have shopping list of the same size. In future work, we will examine the model more sophisticated and heterogeneous shopping lists.
- In our simulations, we consider four different types of attraction values:
  - **Homogeneous.** All zones have the same attraction values, so that no part of the store is more appealing to customers than others.
  - **Proportional to number of sales.** We let the attraction value of each zone be proportional to the number of items that were historically sold from that zone.
  - **Proportional to dwell time.** We let the attraction value of each zone be proportional to the total dwell time of that zone.
  - **Proportional to number of visits.** Similarly, we let the attraction value of each zone be proportional to number of visits of that zone.

5. Results

We compare the results of our model to empirical data by calculating journey time, dwell time, and number of visits. We vary the parameters p (probability of detour), $|L|$ (number of zones in shopping list) and ω_i (attraction value of each zone i) and examine the model's behaviour. For each set of parameter values, we simulate 500 homogeneous customers.

Journey time

In Figure 3, we show the distribution of journey times for $p = 0.2$. We fix shopping-list lengths $|L| = 3$ and take homogeneous attraction values ($\omega_i = 1$ for all nodes). We see similar distributions of journey times with other values of $|L|$ and heterogeneous attraction values. Unfortunately, the journey time is not well captured by our model.

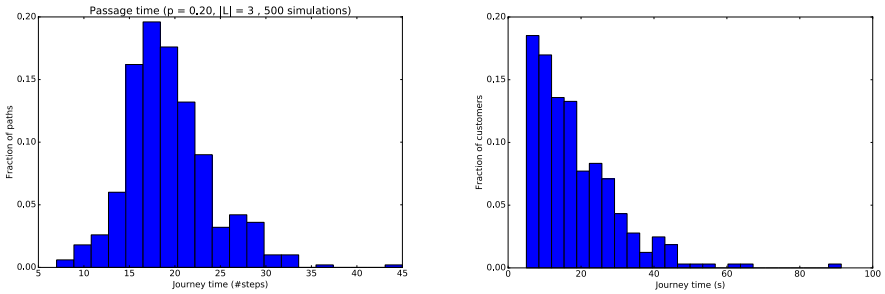

Figure 3: Journey time produced by model (left) and given by data (right). The distribution of the journey time is qualitatively different between the model and the data, so we need to improve our model. For our simulations, we excluded paths less than 5 minutes, so we have excluded these from the plot of the data.

The R^2 value gives a measure of correlation. The highest value is 1 (perfectly correlated) and the lowest is 0 (no correlation).

Dwell time

We compare the predicted dwell time in the zones with the data for different parameters values of p and ω. We calculate R^2 values for the different parameter values (see Figure 4 (left)) and obtain the best fit using either actual dwell time as attraction values or homogeneous attraction values and high p (probability of detour). However, even with the parameter values which give the highest R^2 values, we do not see particularly high correlation (see Figure 5 (right)).

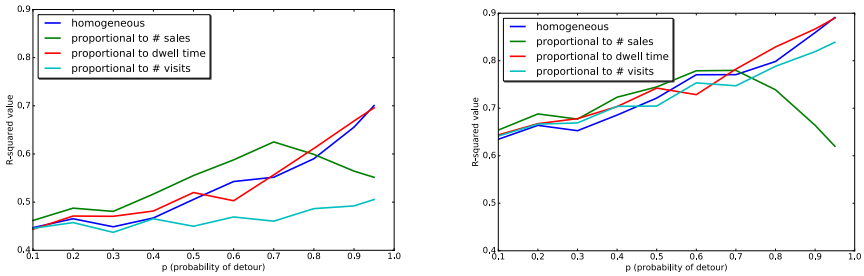

Figure 4: R^2 statistic for dwell time (left) and number of visits (right) between model and data for $L = 3$ and different p (probability of detour) and attraction values. Higher values indicate better fit. For dwell time, we see none of the parameter values give a good fit. For the number of visits, both homogeneous attraction and empirical number of visits give a good fit.

The left hand scatter plot in Figure 5 reveals one of the main reasons why the dwell time is not well captured. Zones near the entrance (1–3) and zones near the tills (49–53) naturally receive many visits, since customer journeys start and end there. In our model, the time that a customer spends at each visit of a zone (also called *waiting time*) is fairly homogeneous, whereas it is rather heterogeneous in the data (see Figure 6). From the data we see that customers spend on average less time in zones 1, 2, 49, 50 and 52 than other nodes, leading to an overestimation of dwell time in our model for these zones (see Figure 5 (left), top left hand corner).

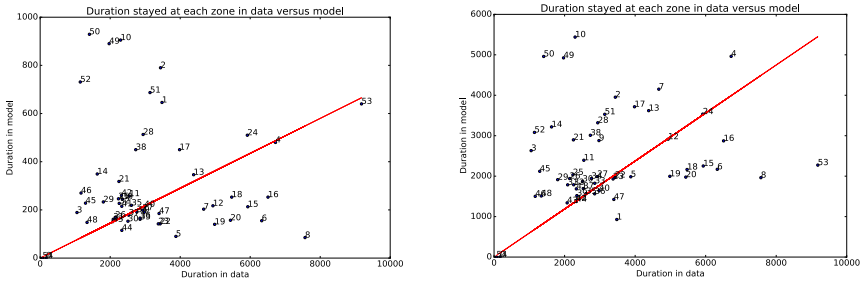

Figure 5: Scatter plots of zone dwell times between model (vertical axis) and data (horizontal axis) for (left) $p = 0.5$ and homogeneous attraction values and for (right) $p = 0.95$ and attraction values equal to the actual dwell time from the data. The red line indicates the line of best fit (through origin). The right figure gives the highest R^2, but is still a poor fit.

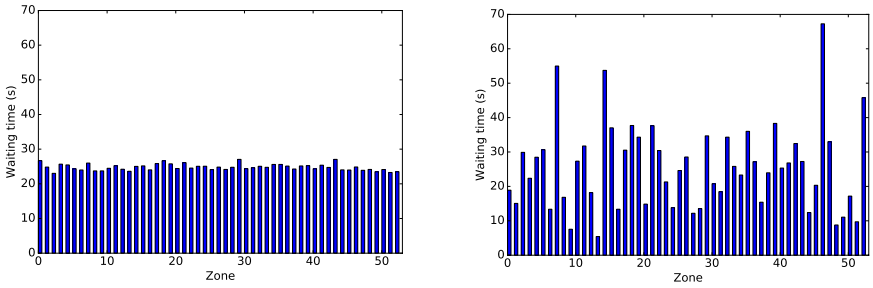

Figure 6: Waiting time of each zone in (left) model and (right) data. In our model, the waiting times are homogeneous, whereas in the data, it is very heterogeneous.

Our model gives a reasonable agreement between model and data for number of visits of each zone, but it does poorly for passage time and dwell time.

Visits

We compare the number of visits between our model and the data as for dwell time (see Figure 4 (right) and 7). When attraction values are homogeneous or proportional to the number of sales, we see large R^2 values at high p (probability of detour). This agreement is further confirmed in the scatter plots (see Figure 7). The best fit is found using homogeneous attraction values and $p = 1$. This corresponds to a pure random walk. With these parameters our model is able to give a reasonable prediction for the number of visits to each zone. Our model is thus able to capture some of the non-temporal statistics reasonably well.

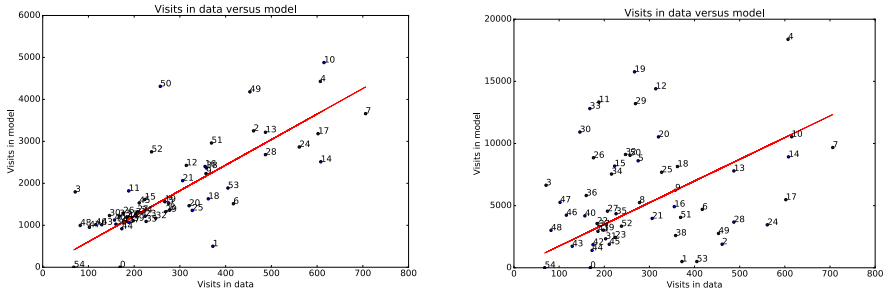

Figure 7: Scatter plots of number of visits between model (vertical axis) and data (horizontal axis) with (left) homogeneous attraction values and (right) with attraction values proportional to the number of items sold from the data (p=0.95 for both). The red line indicates the line of best fit (through origin). Using homogeneous attraction values give a reasonable match, unlike using the number of items sold.

Extension and variations

Our model captures the total number of visits to a given zone, summed over all customers, reasonably well, but fails to accurately capture the temporal dynamics (total dwell time and passage time).

There are many ways to improve and generalize our model to make it more realistic. This includes both incorporation of insights from analysis of the Tesco data and direct incorporation of empirical data itself. We list a few possibilities below.

- **Generation of shopping list L based on sales data.** At the moment we have shopping lists of the same size and with their zones chosen uniformly at random. We can use empirical data such as actual receipts to generate the more realistic shopping lists.

- **Heterogeneous attraction values.** It is likely that different customers have different preferences, and therefore attraction values. Based on customer profiles (e.g. families, students), we could generate a distribution of attraction values. Another way to introduce heterogeneity is to have attraction values based on the items in the current shopping basket. One could imagine that zones which contain items that are frequently bought together with items in the shopping basket receive higher attraction value.

- **Heterogeneous waiting time.** It is likely that customers spend more time in zones where they have something on their shopping list. We could incorporate this by allowing customer to make additional steps in these zones.

6. Conclusion and Future Work

In this mini-project, we have developed a preliminary way of identifying paths from the data and proposed a random-walk-based model for customer journeys. From our simulations, we observed that our model does not capture the temporal statistics (passage time and dwell time) well, but it is able to reproduce the number of visits to each zone reasonably well. Further work will need to be done in order to be able to predict temporal dynamics (such as congestion).

The next stages will be to develop a more rigorous way of identifying paths by conducting a store visit and observing the use of trolleys by customer and staff ourselves. We rely (for our analysis) on the observed behaviour, so it is important to have certainty in identification of customer paths. We will also examine the temporal dynamics in more detail, as our model fails to capture this at the moment. In particular, we will look into the distribution of waiting times and try to incorporate this into our model (e.g. using queueing theory).

7. Potential Impact

In the short-term, the mathematical model in this report provides a framework to model customer journeys. Essential features (e.g. temporal dynamics) are still missing and further work needs to be done, but in the long-term, a more sophisticated model may be able to forecast congestion within supermarkets, which will be useful to analyse store layout changes cost-effectively.

Jeremy Bradley, Lead Data Scientist at Tesco commented on this project that: *'We're very excited about this project; it has the potential to give us a view on how customers are likely to move around our store environment. This is vital for designing better store layouts, helping customers find their products quicker and keeping aisle congestion down. There is real potential for impact with Fabian's work. We look forward to the next stage.'*